Lexical Methods for Managing Variation in Biomedical Terminologies

Alexa T. McCray, Suresh Srinivasan, Allen C. Browne

National Library of Medicine
Bethesda, Maryland

ABSTRACT

Access to biomedical terminologies is hampered by the high degree of variability inherent in natural language terms and in the terminologies themselves. The lexicon, lexical programs, databases, and indexes included with the 1994 release of the UMLS® Knowledge Sources are designed to help users manage this variability. We describe these resources and illustrate their flexibility and usefulness in providing enhanced access to data in the UMLS Metathesaurus®.

INTRODUCTION

There is a great deal of lexical variation in the vocabulary of a natural language. This variation may be rule-governed, or it may be quite idiosyncratic. The variation may be morphological, that is, it relates different forms of the same lexical item through inflection or derivation, or the variation may be simply orthographic, that is, it relates different spellings of the same lexical item. Morphological variation is fairly well understood and is described in several standard references (e.g., [1-3]), and orthographic variation is generally studied either from the point of view of spelling errors or from the point of view of variant spellings in particular dialects [4-6].

The development of methods for capturing lexical variation in computerized systems is, however, a difficult problem because of the wide range of possible variations and the possibility for seemingly unconstrained combinations of these variations. The development of so-called stemming algorithms and spelling error detection algorithms has been the subject of some research (see, for example, [7-8]).

The availability of the UMLS knowledge sources [9], and especially its Metathesaurus, has led to a number of experiments involving automated lexical matching methods, either as part of the development process [10], or for the purpose of comparing the Metathesaurus content with some other vocabulary (e.g., [11-12]), or in order to identify Metathesaurus concepts in free text (e.g., [13]). Each of these experiments has used lexical methods that have some similarity to the others, but that also differ in a variety of ways.

The 1994 release of the UMLS knowledge sources includes a fourth knowledge source, the SPECIALIST™ lexicon, together with a set of lexical programs. The lexicon has been developed in the context of the authors' work in biomedical language processing [14]. The lexical programs generate a range of variations for English lexical items and should be useful for recognizing and thereby abstracting away from lexical variation in biomedical terminologies and texts.

SPECIALIST LEXICON

The SPECIALIST lexicon is an English language lexicon containing many biomedical terms. The lexicon entry for each word or term records syntactic, morphological, and orthographic information. Lexical entries may be single or multi-word terms. Entries which share their base form and spelling variants, if any, are collected into a single lexical record. The base form is the uninflected form of the lexical item; the singular form in the case of a noun, the infinitive form in the case of a verb, and the positive form in the case of an adjective or adverb. Currently the lexicon contains some 60,000 records, with approximately 120,000 forms. Lexical information includes syntactic category, inflectional variation (e.g., singular and plural for nouns, the conjugations of verbs, the positive, comparative, and superlative for adjectives and adverbs), and allowable complementation patterns (i.e., the objects and other arguments that verbs, nouns, and adjectives can take).

Lexical items are selected for coding from a variety of sources. These include data from MEDLINE® citation records, terms in the Dorland's Illustrated Medical dictionary, the 10,000 most frequent words listed in the American Heritage Word Frequency book and the 2,000 lexical items used in the controlled definitions of Longman's Dictionary of Contemporary

0195-4210/94/$5.00

English. Lexical records are created using a lexicon building tool called Lextool. Lextool is a menu based system that accepts as input either a file of lexical items or lexical items entered at the keyboard. Lextool is supported by an underlying lexical grammar that constrains the information that can be entered for lexical items of a particular syntactic category and also serves to validate the information that has been encoded. A variety of reference sources is used in coding the lexical records. These include dictionaries of general English (primarily learner's dictionaries), medical dictionaries, and data derived from actual usage of the lexical items in MEDLINE.

The SPECIALIST lexicon is distributed in both unit record and relational table format. The unit record format is a frame structure consisting of slots and fillers. The slots are the basic lexical attributes, such as syntactic category, variants, complements, etc. The fillers express the possible values of those attributes for that particular lexical item. The relational table format expresses the same information in ten tables. These tables have been created so as to maximize their usefulness for different types of applications. For example, there is one table that contains only agreement and inflection information, another for complementation patterns, a table for spelling variants, and another table for abbreviations and acronyms and their fully expanded forms.

The lexicon is also available for lookup and browsing on a World Wide Web server. The URL for the server is http://wwwetb.nlm.nih.gov/. The lexicon can be found under the Information Technologies/Natural Language Systems Program menu items.

LEXICAL PROGRAMS

The lexical variant programs are written in C and use data from the SPECIALIST lexicon as they compute the different forms of lexical items. The lexical programs consist of several different modules that can be combined in a variety of ways to generate variants. For example, users may be interested in seeing only the singular or plural of an input term. In that case, they would choose the inflection option. Or, they may be interested in running their terms against a stop word list and also ignoring word order so as to maximize their chances of finding related terms in a particular vocabulary or text. In this case, they would choose the stopword removal and word sorting options.

The programs allow for a good deal of flexibility in matching one term to another. The basic principle that is involved in using the programs is that any manipulation of a (source) input term or terms must involve the same manipulation of the target terms. For example, if users want to see if terms in their particular vocabulary (source) are found in the Metathesaurus (target), and if they want to find those terms regardless of whether they appear there in the singular or plural or whether they are in upper or lower case, then they would transform the source vocabulary using the lowercasing and inflectional options and, importantly, they would do the same for the Metathesaurus terminology.

Normalization Routines

Since some users will prefer to use a method that does not involve additional processing of the Metathesaurus data, a normalization program ("norm") together with a normalized string index of all Metathesaurus terminology is included with the UMLS Knowledge Sources. The norm program is essentially one set of lexical variant options. The normalization process involves splitting a string into its constituent words, lower-casing each word, converting each word to its base form, ignoring punctuation, and sorting the words in a multi-word term into alphabetic order. This means that when matching a normalized string in a source vocabulary to the Metathesaurus normalized string index, alphabetic case, inflectional variation, punctuation, and word order are ignored in the comparison.

Some examples of the normalization process using the 1994 Metathesaurus terminology as the target vocabulary and a list of terms provided by J. Vries as the source vocabulary are discussed below. The source vocabulary terms are from the University of Pittsburgh MARS system (see [15] for some discussion of this system), and originally did not readily map to Metathesaurus terms. The examples serve to illustrate the normalization process.

The source term "abdominal binder" does not exist in that exact form in the Metathesaurus. The two forms "Binders, Abdominal" and "Abdominal Binders" do, however, exist. The normalized index entry for both of these forms is "abdominal binder", and, thus, the term is found. The source term "battery" maps to the term "Batteries" in the Metathesaurus because "battery" is the normalized form of the Metathesaurus term. And, "eye-patch" maps to the terms "Patches, Eye", and "Eye Patches" because all are normalized to

"eye patch". In these cases, the general English inflectional pluralization rules (add "s", convert "y" to "ies", and add "es" to words ending in "ch, sh, s, or x") have been recognized as part of the normalization process. In addition, word order and alphabetic case have been ignored.

The following two examples illustrate that the normalization routines recognize Greco-Latin inflectional variation. The source term "nasal cannula" is not found in the Metathesaurus, but its plural "Cannulae, Nasal" is found and is mapped through normalization to "nasal cannula". The source term "elbow prosthesis" maps to "Prostheses, Elbow" through normalization. Irregular plurals are also handled by the normalizer, if the information is stored in the SPECIALIST lexicon. This allows the source term "gamma knife" to map to the Metathesaurus term "Gamma Knives". The source term "blood type" is also not found in the Metathesaurus, but its inflectional variant "Blood Typing" is found through normalization.

Some examples of word order variation are shown in the next two examples. As noted above, the normalization routines ignore word order in multi-word terms, since they alphabetize the words in source and target terms. The normalized form of the source term "introducer catheter" is "catheter introducer", as are the normalized forms of the Metathesaurus terms to which it maps: "Catheter Introducers" and Introducers, Catheter". Ignoring word order may in some cases lead to the well-known "venetian blind", "blind venetian" phenomenon. That is, if two terms that vary only with regard to word order and that have different meanings do exist in the target vocabularly, both will be retrieved through the normalization routines. When using the normalization routines in unconstrained contexts, such as free text, it is wise to review the results for cases where concepts might fall together. Review of the Metathesaurus terminology, however, has yielded very few of these types of examples. Since word order is highly variable in the Metathesaurus vocabularies, there appears to be significant benefit in abstracting away from it.

The source term "meckel diverticulectomy" does not exist in that form in the Metathesaurus, but two genitive (possessive) forms do: "Diverticulectomy, Meckel's" and "Meckel's diverticulectomy". In this case, ignoring punctuation and ignoring sequences smaller than two characters (i.e., "s") give the normalized form "diverticulectomy meckel" for both source and target terms. The option "remove genitive markers" of the lexical variant generation programs, which will be discussed below, would give the same result.

With the 1993 release of the UMLS knowledge sources, an index of all the words in the Metathesaurus was provided. The 1994 release again contains a word index, and it also contains a normalized word index in which all the words have been normalized according to the routines discussed above. In some cases, use of the normalized word index will provide additional terminology of interest. For example, if the user were interested in finding all terms in the Metathesaurus that include the word "suture", using the simple word index would yield the following, among others: "Suture", "Closure by suture", "Cranial Suture", "Suture granuloma", "Suture Technic", and "suture line care". The use of the normalized word index would yield all those terms as well as terms such as the following: "Suturing" and "Congenital ossification of sutures", since "suturing" and "sutures" are both normalized to "suture".

Lexical Variant Generation

In some cases, the normalization routines may not give the desired results. This may be because the source vocabulary or text has certain characteristics that are not accounted for by normalization, or it may be because the user would like to be more "aggressive" in the matching routines (that is, by accepting greater variation, with the hope that there will be some correct matches). In this case, the user may decide to use some of the other options that are provided as part of the lexical variant generation (lvg) programs. For example, the lvg stopword option removes highly frequent common words such as "of, and, with, for, in, by", etc. Using this option together with the word order option would, for example, match the term "splenic artery aneurysm" to the Metathesaurus term "Aneurysm of splenic artery".

Use of the lvg derivational morphology module allows the user to find closely related terms that may not have the same syntactic category, but that are usefully related nonetheless. For example, if the source vocabulary or text includes the adjective "hyperplastic", using the derivational option will map this to the noun "Hyperplasia", which is a Metathesaurus term. Nominalizations (noun forms of verbs or adjectives) are prevalent in the biomedical vocabulary. Often a medical dictionary will list only the nominalization and will not list the verb or adjective form. When mapping terms from free text to the Metathesaurus, it might prove fruitful to use the derivational module of

lvg to identify such variants. For example, verbs such as "aspirate", "consume", and "deceive" would map to the nominalized Metathesaurus terms "Aspiration", "Consumption", and "Deception", respectively. Analogously, the adjectives "bacterial", "endometrial", and "ganglial" would map to the Metathesaurus nouns "Bacteria", "Endometrium", and "Ganglia", respectively.

The morphology modules of lvg are based on a rule and fact paradigm designed to capture the morphological relations between terms. Rather than analyzing words into morphemes and describing morphological relations in terms of morphemes and their meanings, the program captures common morphological relations. Derivational and inflectional morphology are both handled by a set of rules (with any exceptions noted) and associated facts. Derivational morphology deals with the alternations between lexical items that often involve a change in syntactic category, or part of speech. For example, "malaria" and "malarial" are related through derivational morphology. "Malarial" is the adjectival form of the noun "malaria". This relationship is captured in the form of a heuristic rule stating that nouns ending in "-a" often correspond to adjectives ending in "-al". Rules are recorded in a relational format of the form: "suffix 1| syntactic category 1| suffix 2| syntactic category 2". This rule states that a term of syntactic category 1 ending in suffix 1, may be morphologically related to another term of syntactic category 2 ending in suffix 2. The rule for "malarial" and "malaria" has the form: "al|adj|a|noun". Rules are symmetric, e.g., "al|adj|a|noun" is equivalent to "a|noun|al|adj". Derivational variation is rule-governed to some extent, but some alternations are more productive than others. The effectiveness of these rules is increased by recording for each rule a list of known exceptions. For example, "aura" and "aural" are not related (they mean different things) and are, therefore listed as known exceptions to the rule "al|adj|a|noun". Exceptions to rules have been discovered empirically by comparing words from various machine readable sources, including the Unix system dictionary, Dorland's Illustrated Medical Dictionary, and the Oxford Advanced Learner's Dictionary.

Not all instances of derivationally related words are productive enough to be usefully stated as rules. Particular instances of morphologically related words are recorded as facts in a similar format to the rules. Examples of facts used by the derivational module are: the adjective "presidential" related to the noun "president", the adjective "tyrosinate" related to the noun "tyrosine", and the noun "column" related to the adjective "columnar".

The inflectional rules and facts are similar to the derivational rules and facts with appropriate changes. For example, nouns ending in "us" often have plurals in "i" as in "focus" and "foci". This inflectional fact is also recorded in terms of a heuristic rule stating that singular nouns ending in "us" may have plurals ending in "i". This rule is like the derivational rules discussed above except that an additional field indicates the inflection that the suffix signals. Most of the inflectional rules are derived from the inflectional classes used by the SPECIALIST lexicon. The rule just mentioned is part of the Greco-Latin (glreg) inflectional class in the lexicon. Just as with derivational rules, known exceptions may be listed with the rule.

LEXICAL DATABASES

Three databases that may be useful for some developers have also been provided. The first (dm.db) is a file that contains some 10,000 pairs of known derivational variants. The rules and facts used by the derivational morphology module have been drawn from this file. The file relates pairs of words that are derivationally related and gives their syntactic categories. Sample terms that are listed there are:

pharyngeal (adj)|pharynx (noun)
disabled (adj)|disability (noun)
comply (verb)|compliance (noun)
blastogenic (adj)|blastogenesis (noun)
transparent (adj)|transparency (noun)
dosage (noun)|dose (noun)

A second database of closely related terms that mean the same thing, but may sometimes differ in syntactic category is provided in the sm.db file. These closely related terms, currently approximately 2,500 pairs, have been drawn from a variety of sources including medical dictionaries and may or may not be represented in the Metathesaurus. If one of the terms in the pair is in the Metathesaurus, but the other is not, then this file may serve to provide additional entry points into that knowledge source. Some examples from the file are:

false paralysis (noun)|pseudoparalysis (noun)
asphyxiation (noun)|suffocation (noun)
ablate (verb)|remove (verb)
pneumal (adj)|lungs (noun)

hepatocellular (adj)|liver cells (noun)
nasal (adj)|nose (noun)
digital (adj) / finger (noun)

The third database contains about 4,000 pairs of spelling variants. These have been extracted from the SPECIALIST lexicon. These may also serve as additional entry points into the Metathesaurus if one of the items in the pair is in the Metathesaurus, but the other is not. Some examples from the file (sp.db) are:

linoleic acid (noun)|linolic acid (noun)
amebicidal (adj)|amebacidal (adj)
leukocyte (noun)|leucocyte (noun)
haematocrit (noun)|hematocrit (noun)
nanogramme (noun)|nanogram (noun)
fibre (noun)|fiber (noun)

CONCLUSION

The lexical methods described above offer a variety of techniques for the management of lexical variation in biomedical terminologies and texts. The indexes provide standard ways to access the UMLS Metathesaurus, and the lexical programs and databases provide users with the flexibility to design their own access methods. Future releases of the UMLS knowledge sources should involve growth and improvement of these resources, particularly as they are used in a variety of applications.

REFERENCES

1. Bauer L. English Word Formation. Cambridge: Cambridge University Press, 1983; 311 pages.

2. Marchand H. The Categories and Types of Present-Day English Word-Formation. Munich: C.H. Beck, 1969; 545 pages.

3. Quirk R, Greenbaum S, Leech G, Svartvik J. A Comprehensive Grammar of the English Language. London: Longman Group Limited, 1985; 1515-1585.

4. Benson M, Benson E, Ilson R. Lexicographic Description of English. Studies in Language Companion Series, 1986; Volume 14:14-18,169-174.

5. Emery DW. Variant Spellings in Modern American Dictionaries. National Council of Teachers of English, 1973; 130 pages.

6. Dirckx JH. The Language of Medicine, its Evolution, Structure, and Dynamics. Hagerstown: Harper & Row Publishers, 1976; 170 pages.

7. Peterson JL. Computer programs for detecting and correcting spelling errors. Communications of the ACM 1980;23(12):676-687.

8. Porter MF. An algorithm for suffix stripping. Programming 1980;14:130-137.

9. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Methods of Information in Medicine 1993;32:281-91.

10. Sherertz DD, Tuttle MS, Blois MS, Erlbaum MS. Intervocabulary mapping within the UMLS: the role of lexical matching. In: Greenes RA, ed. Proceedings of the 12th annual symposium on computer applications in medical care. Los Angeles: IEEE Computer Society, 1988;201-206.

11. Huff SM, Warner HR. A comparison of Meta-1 and HELP terms: Implications for clinical data. In: Miller RA, ed. Proceedings of the 14th annual symposium on computer applications in medical care. Los Angeles: IEEE Computer Society, 1990;166-169.

12. Cimino JJ. Representation of Clinical Laboratory Terminology in the Unified Medical Language System. In: Clayton PD, ed. Proceedings of the 15th annual symposium on computer applications in medical care. New York: McGraw Hill, 1991;199-203.

13. Miller RA, Gieszczykiewicz FM, Vries JK, Cooper GF. CHARTLINE: Providing bibliographic references relevant to patient charts using the UMLS Metathesaurus knowledge sources. In: Frisse ME, ed. Proceedings of the 16th annual symposium on computer applications in medical care. New York: McGraw Hill, 1992;86-90.

14. McCray AT, Sponsler JL, Brylawski B, Browne AC. The role of lexical knowledge in biomedical text understanding. In: Stead W, ed. Proceedings of the 11th annual symposium on computer applications in medical care. Los Angeles: IEEE Computer Society Press, 1987;103-107.

15. Vries JK, Marshalek B, D'Abarno JC, Yount RJ, Dunner LL. An automated indexing system utilizing semantic net expansion. Computers and Biomedical Research 1992; 25:153-167.